Aula 04 - Classificação

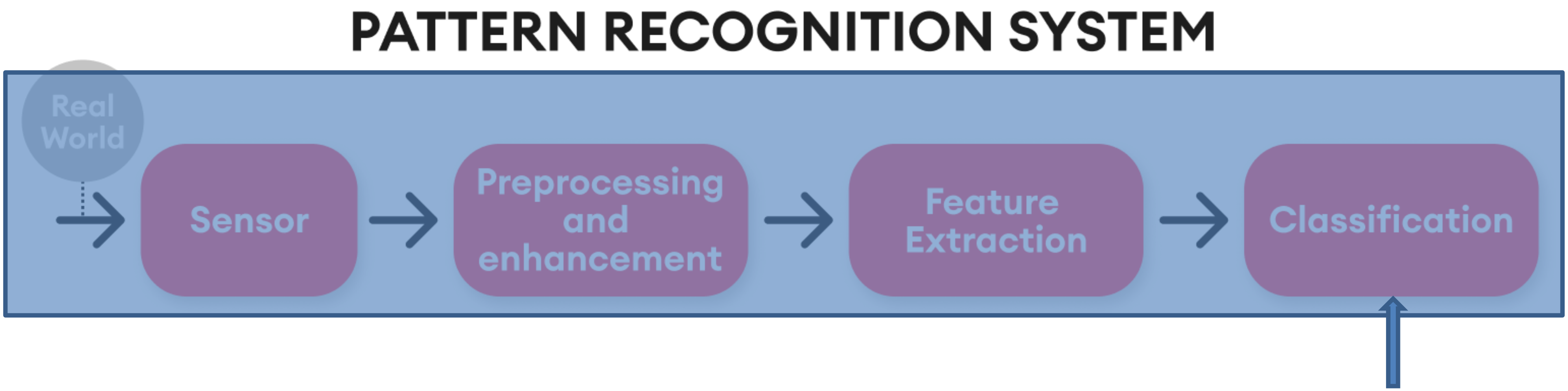
Prof. André Gustavo Hochuli

gustavo.hochuli@pucpr.br
aghochuli@ppgia.pucpr.br

Tópicos

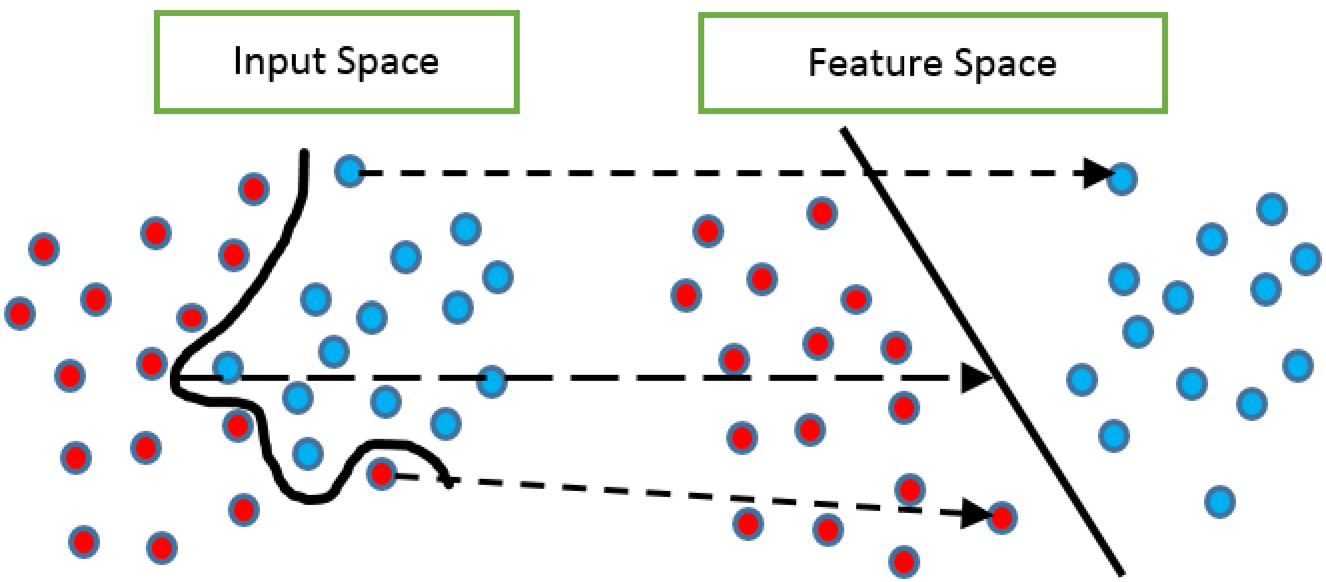
- Discussão Inicial
- Modelos de Classificação
  - K-NN, Logistic Regression, Decision Trees Naïve Bayes, SVM and MLP
- Métricas de Avaliação
  - Accuracy, Precision, Recall and F1-Score
- Practice

Visão Computacional (Workflow)



Problema

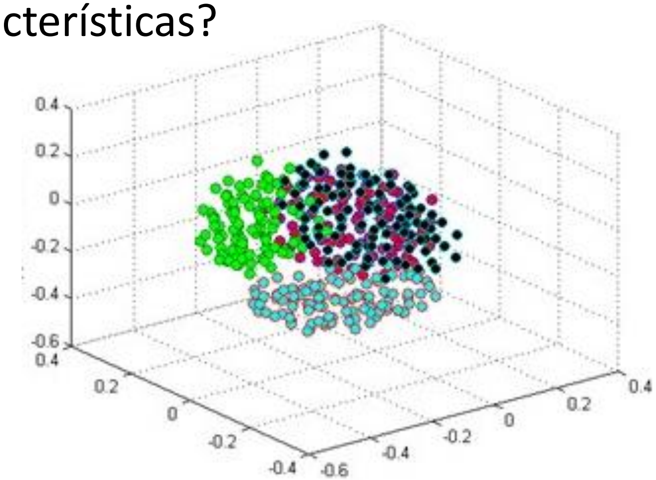
- Até agora temos discutido como extrair características



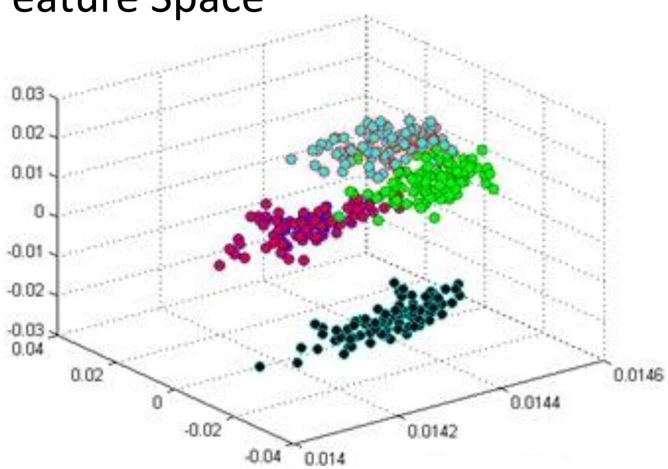
Problema

- Quão discriminante são as características?

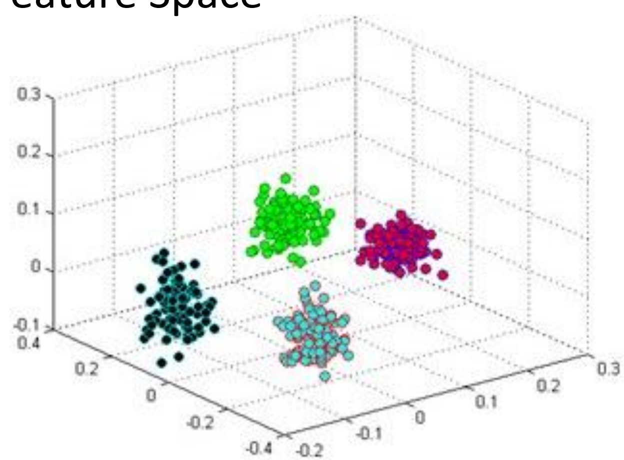
Input Space



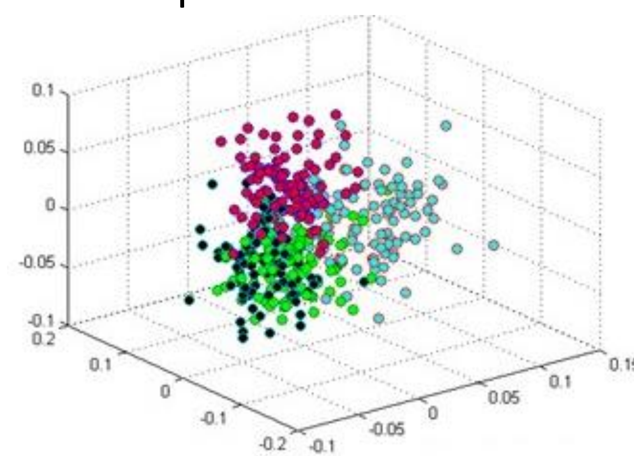
Feature Space’



Feature Space’’

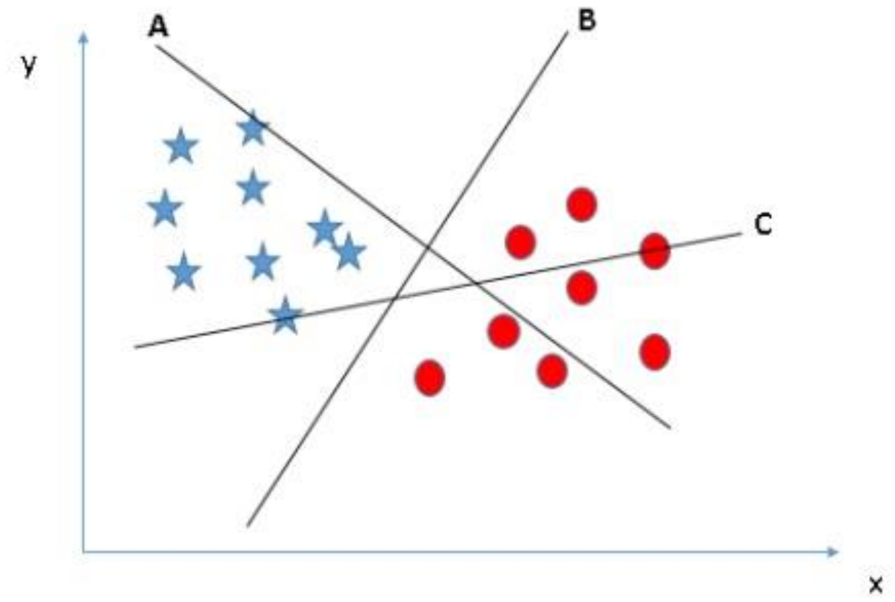


Feature Space ’’’



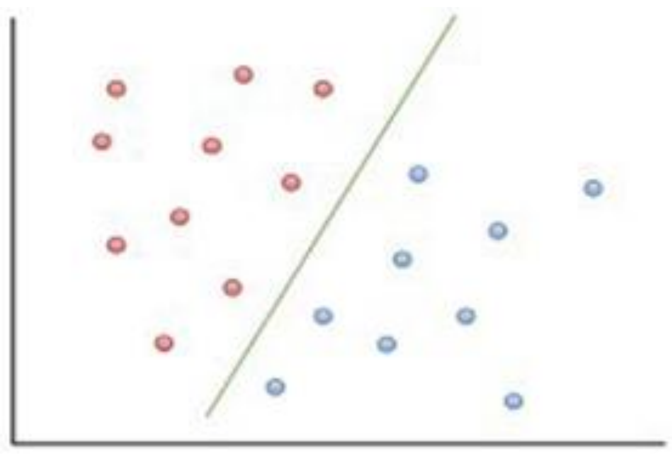
Problema

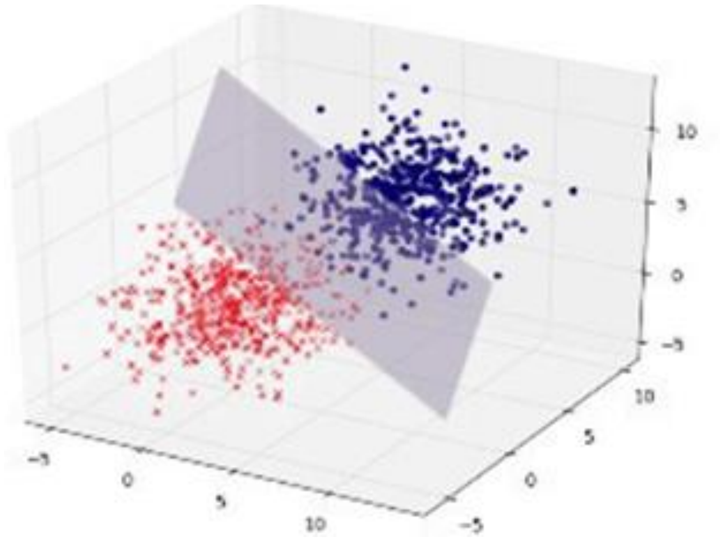
- Como computar a fronteira de decisão?



Problema

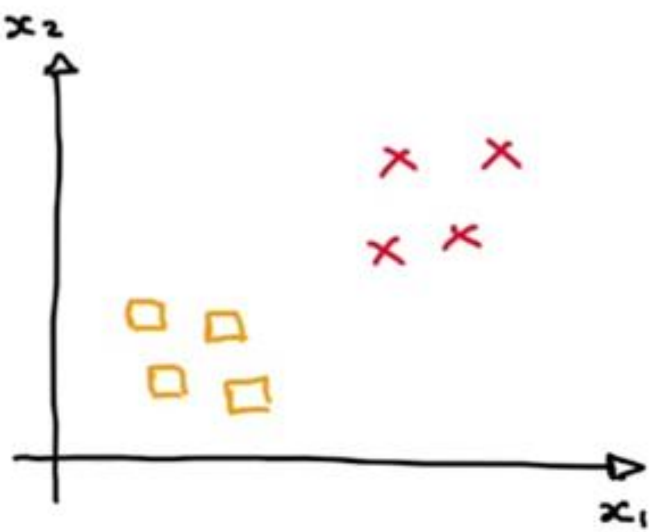
- Hiperplano
  - 2-D, 3-D … N-D (or N-Features)

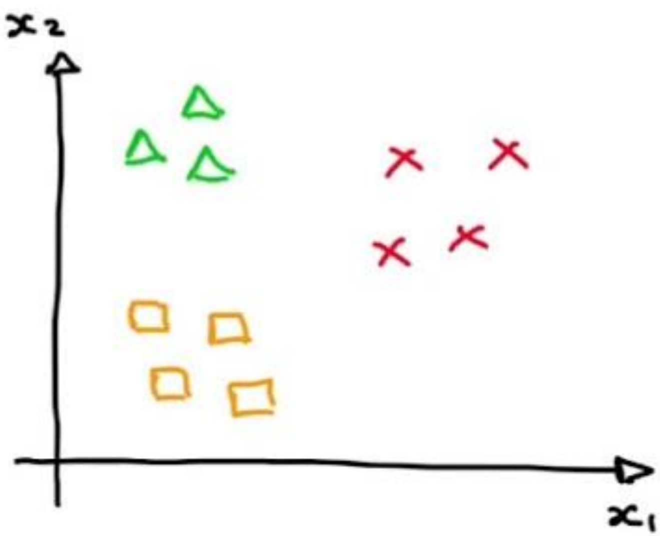




Problema

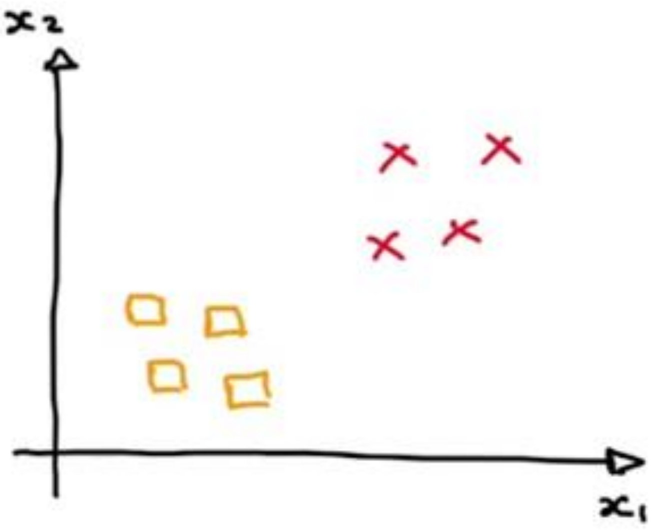
- Classificação Binária vs Multi-Class

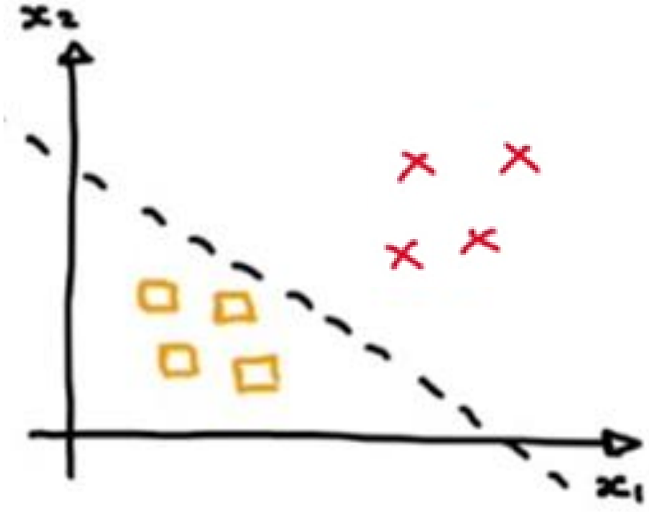




Problema

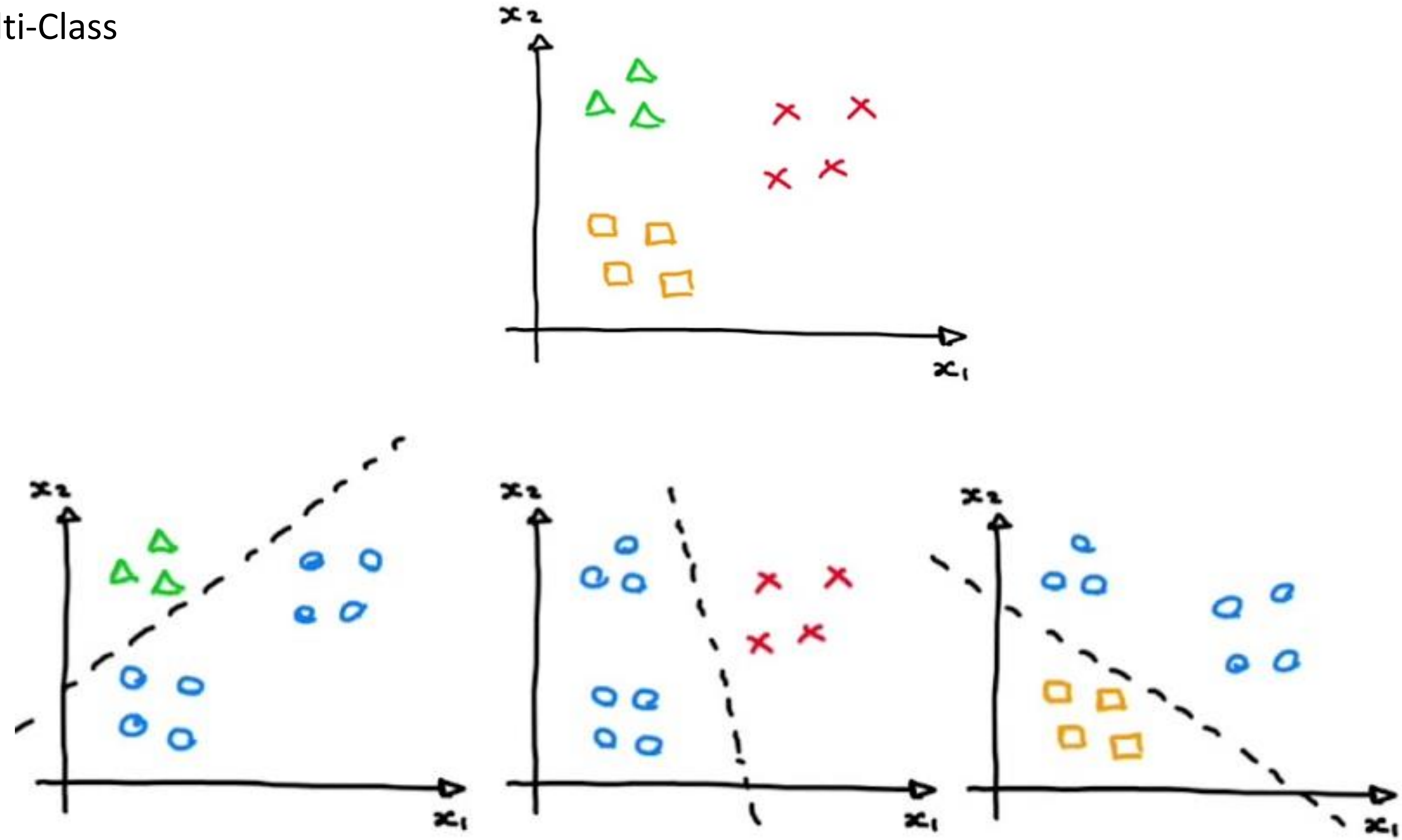
- Classificação Binária





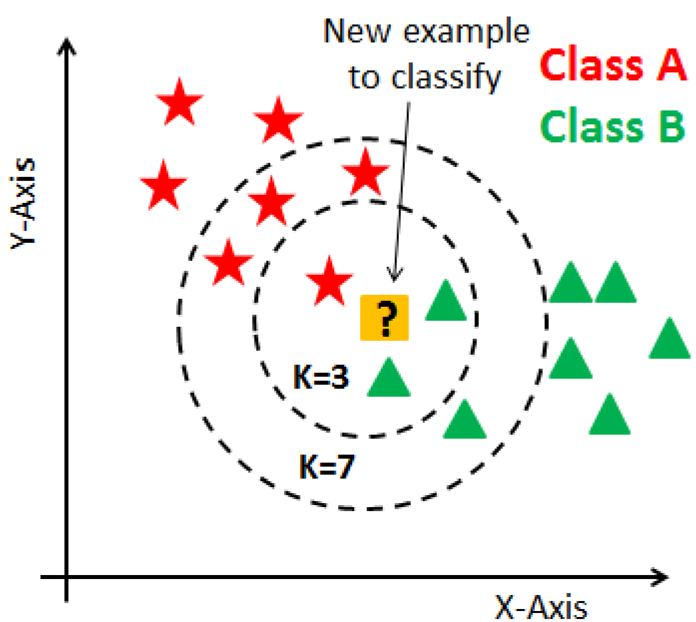
Problema

- Multi-Class



Modelos de Classificação
KNN

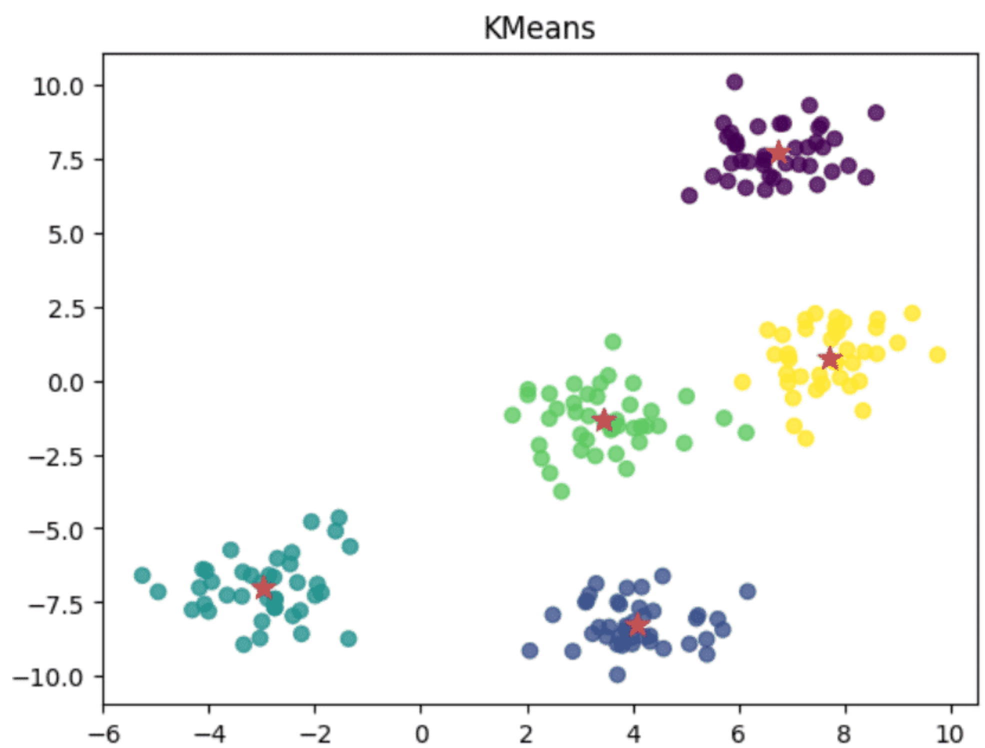
- Computa a similaridade no espaço de característica (Distância Euclidiana, Manhattan….)
- K-Vizinhos mais próximos determinam a classe (Votação)
- Não tem etapa de treinamento. Computa as distâncias para cada amostra de teste



$$d(x,y) = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2}$$

Modelos de Classificação
K-Means

- Calcula a distância entre a amostra de teste e os k-centroides
- Os clusters são definidos na etapa de treinamento

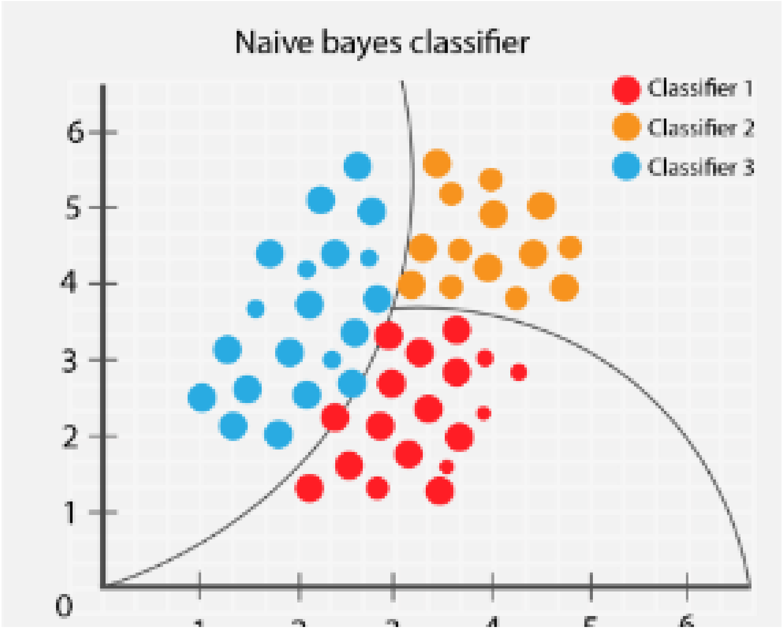


Modelos de Classificação
Naïve Bayes

- Teorema de Bayes
- Probabilidades*: A priori vs Posteriori*

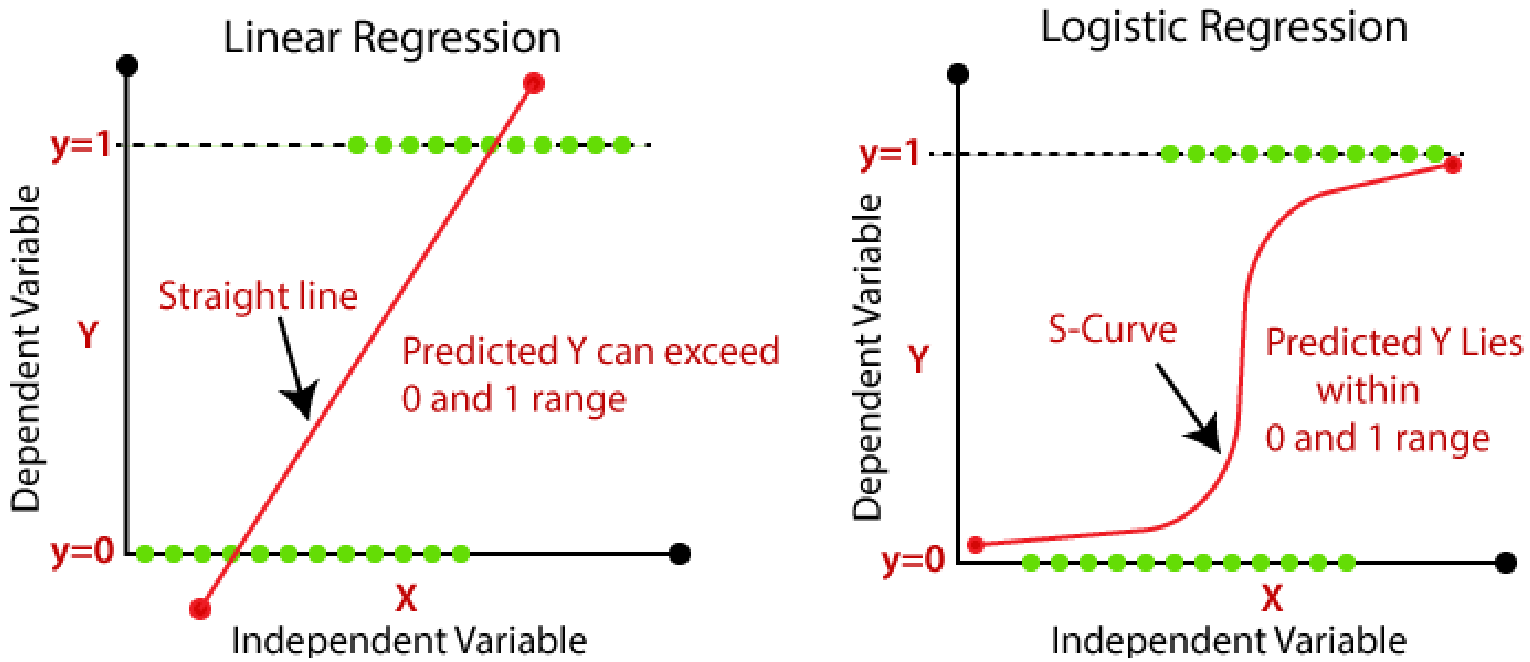
$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

$$\text{Posterior} = \frac{\text{prior x likelihood}}{\text{evidence}}$$



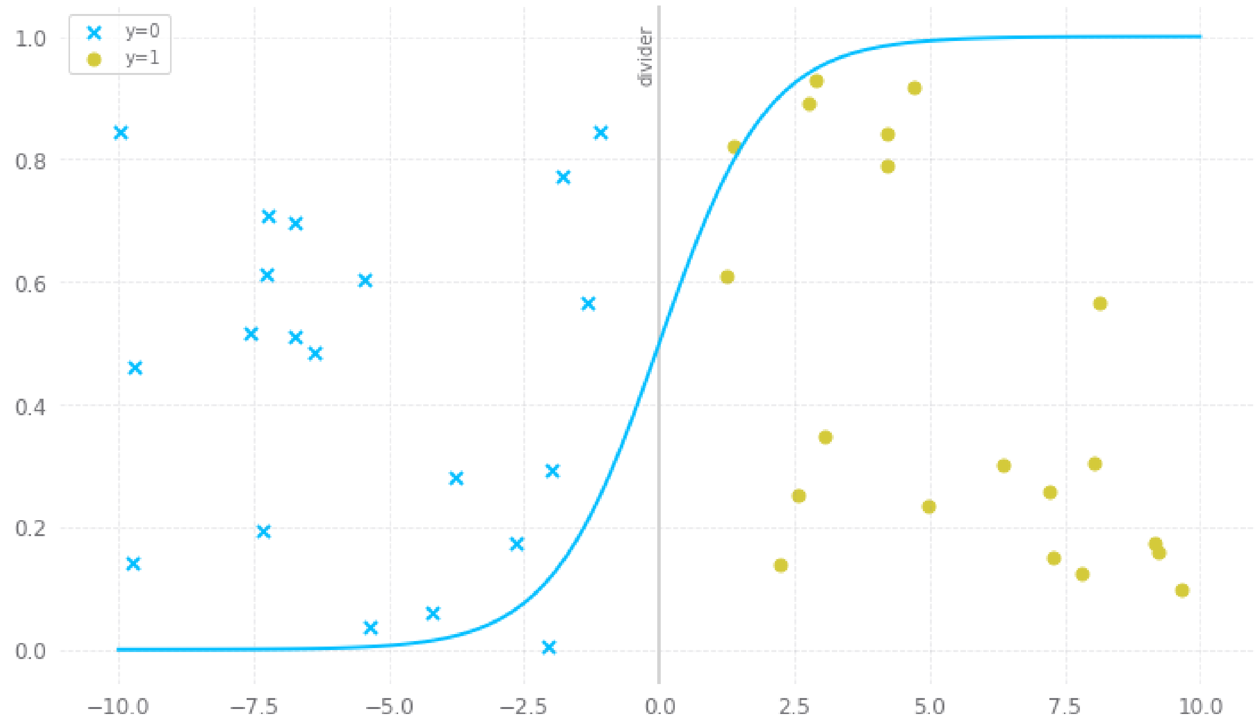
Modelos de Classificação
Logistic Regression (LR)

- Linear vs Logistic



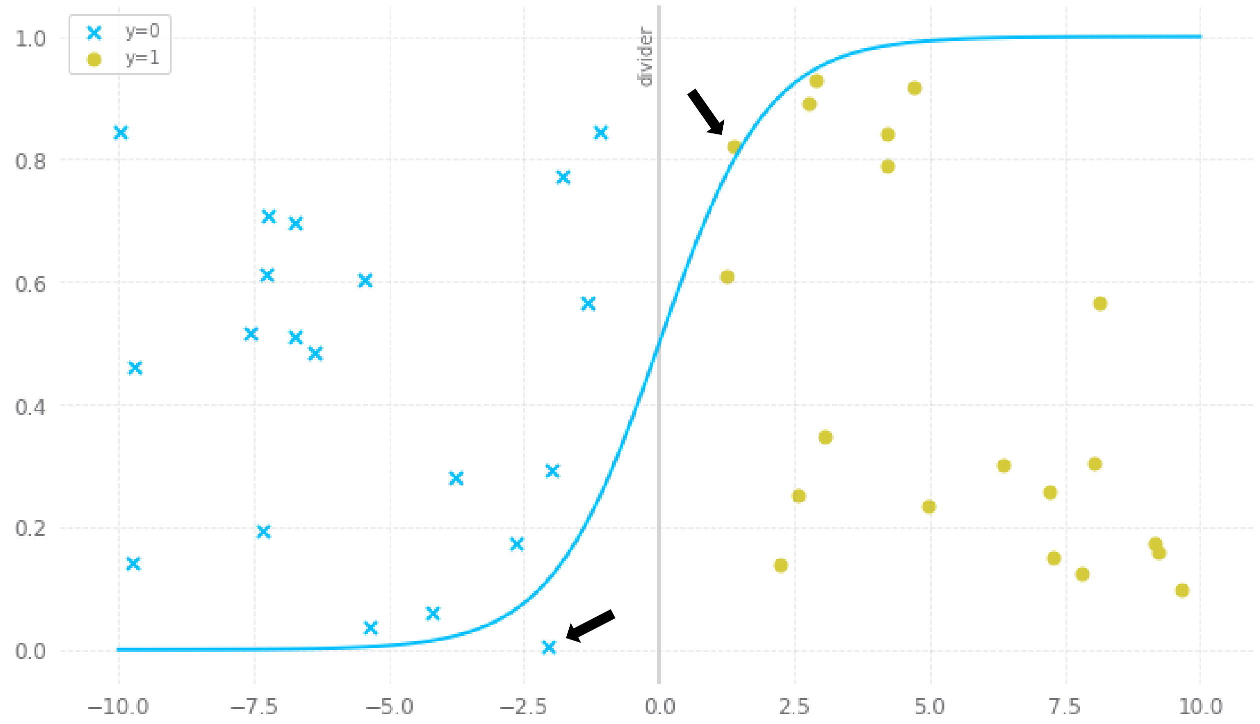
Modelos de Classificação
Logistic Regression (LR)

- Logistic Boundary



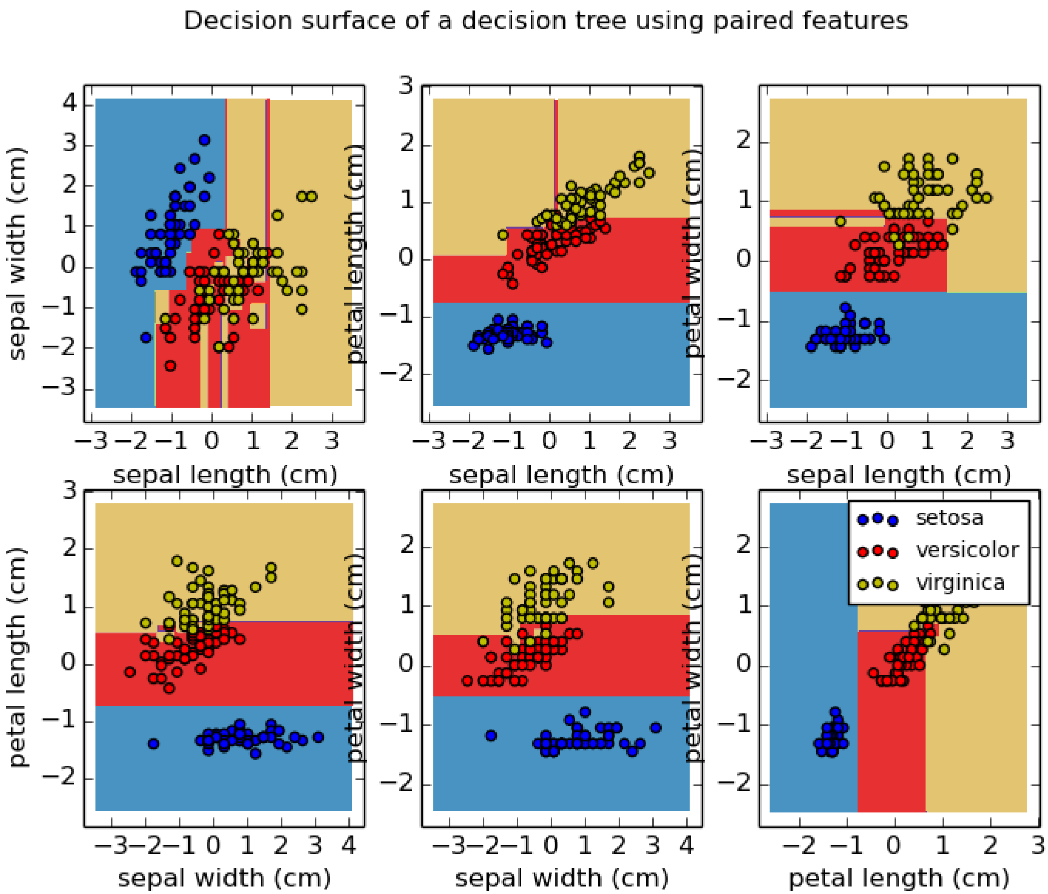
Modelos de Classificação
Logistic Regression (LR)

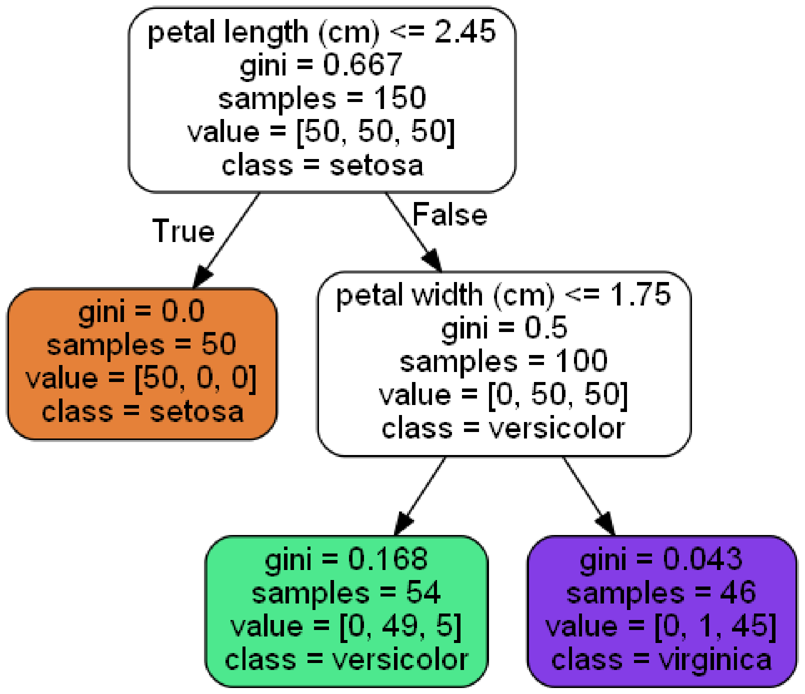
- Logistic Boundary



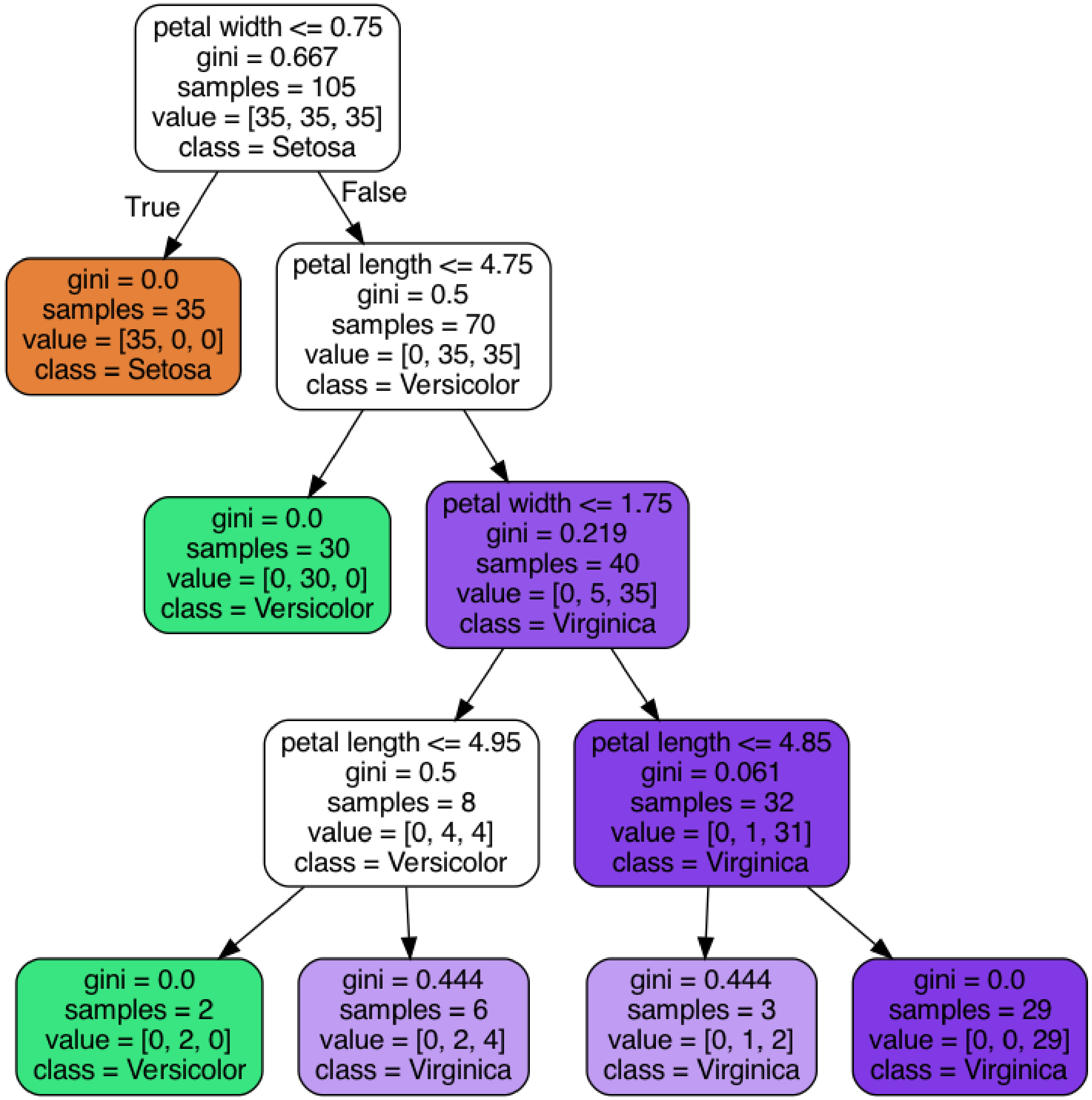
Modelos de Classificação
Decision Tree

- Determina regras de decisão



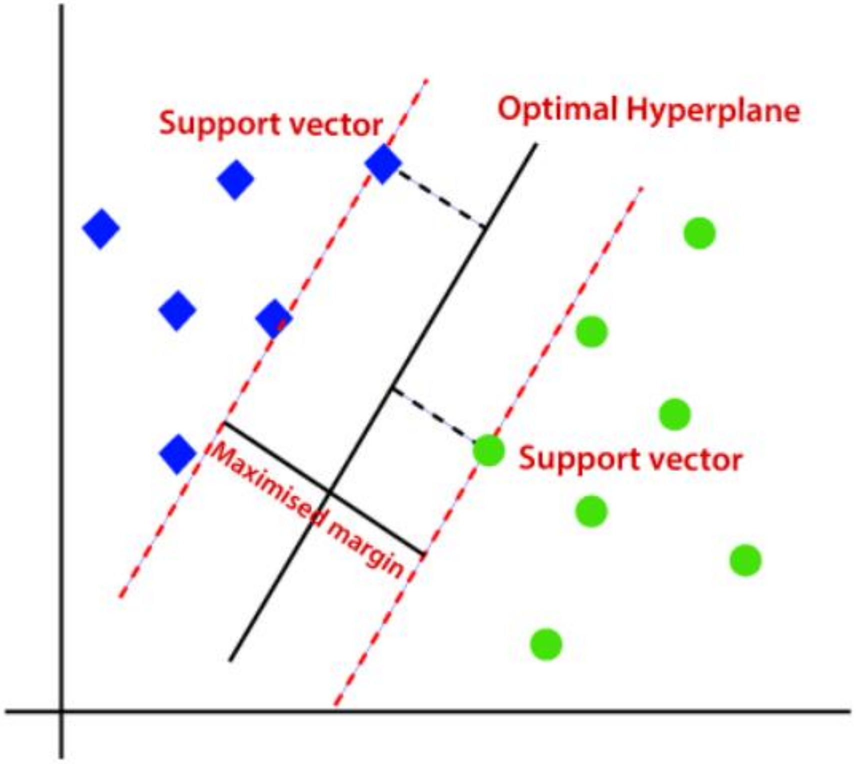


Modelos de Classificação
Decision Tree



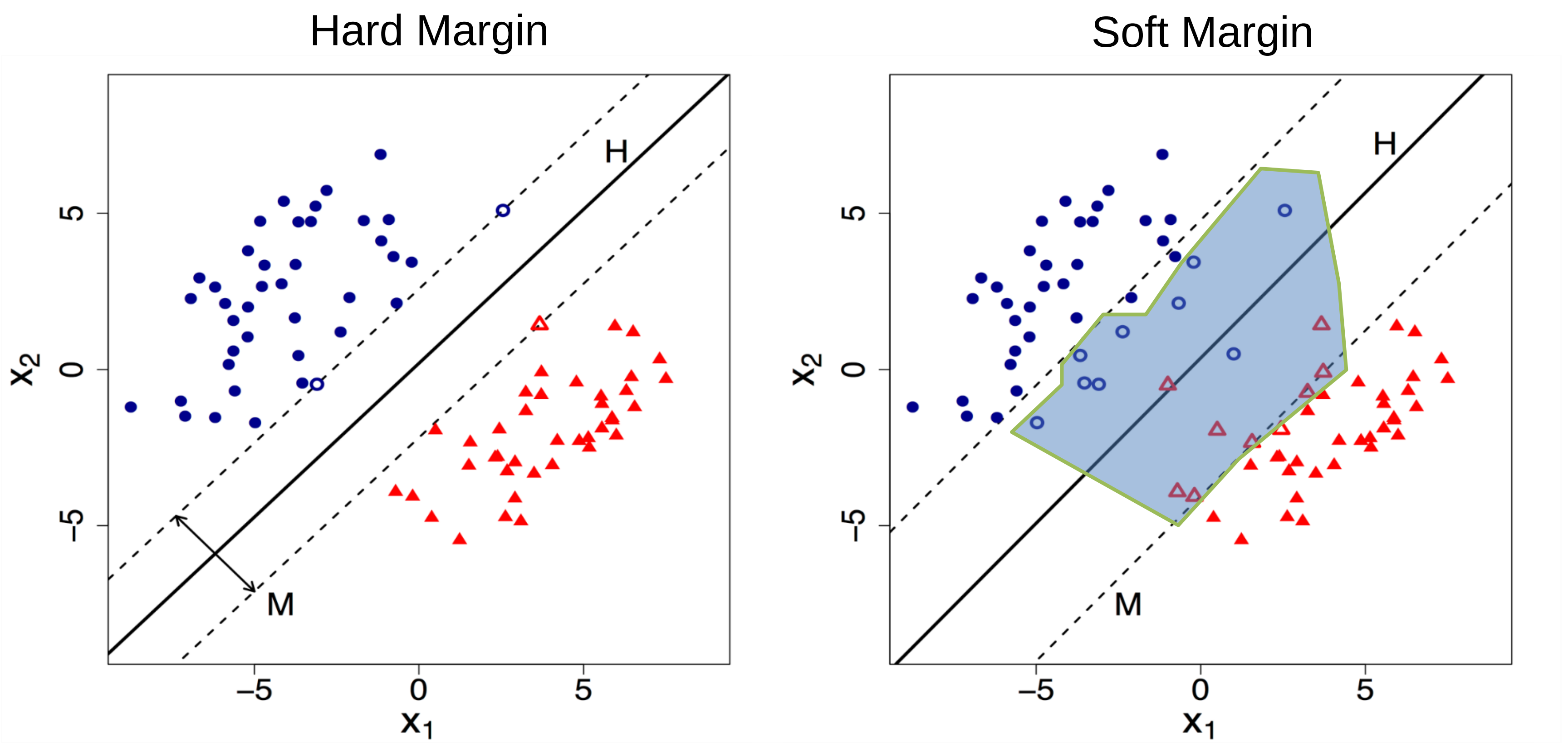
Modelos de Classificação
Support Vector Machine (SVM)

- Fronteiras de decisão são baseadas em vetores de suporte



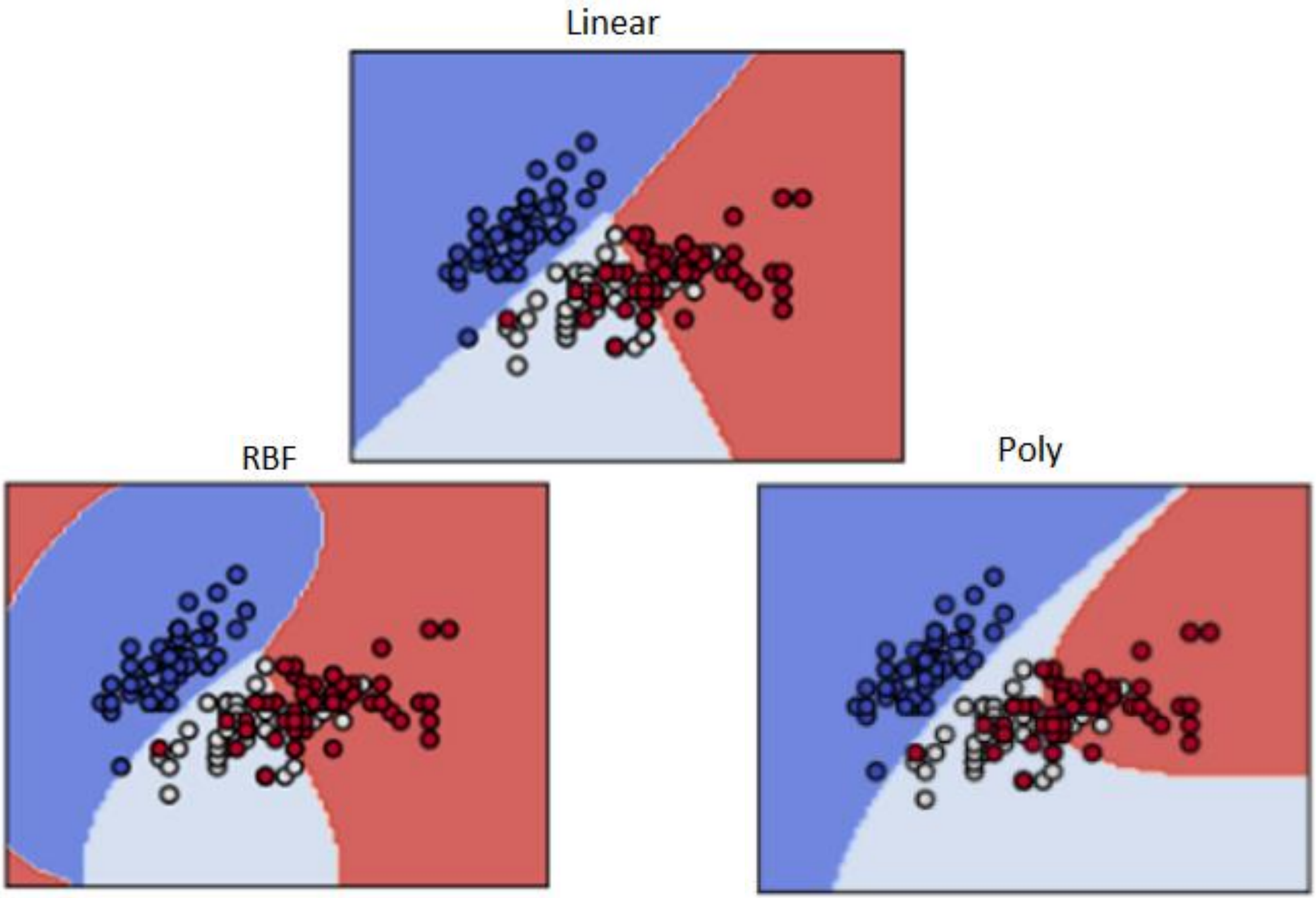
Modelos de Classificação
Support Vector Machine (SVM)

- Fronteiras de decisão são baseadas em vetores de suporte



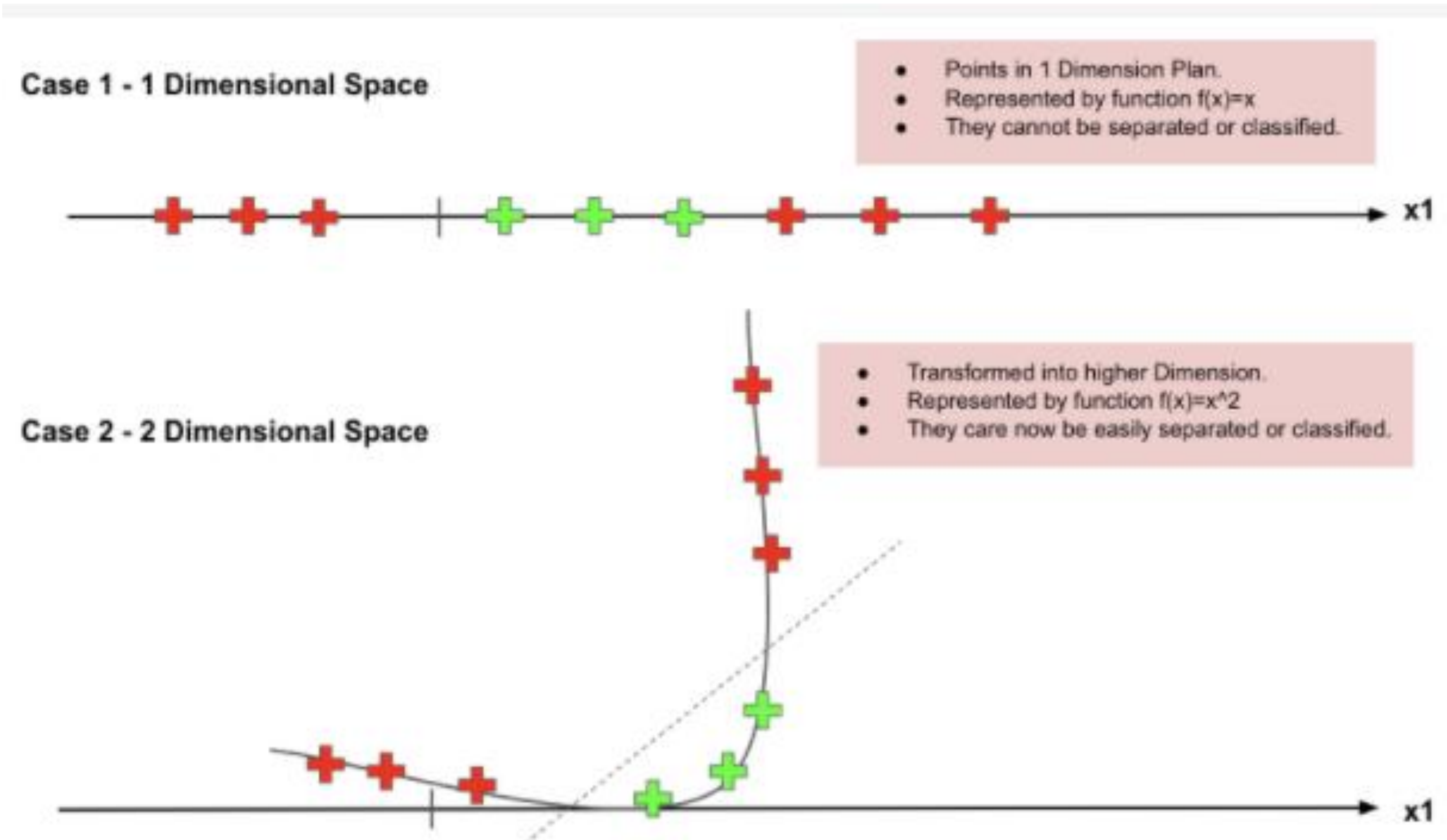
Modelos de Classificação
Support Vector Machine (SVM)

- Kernels



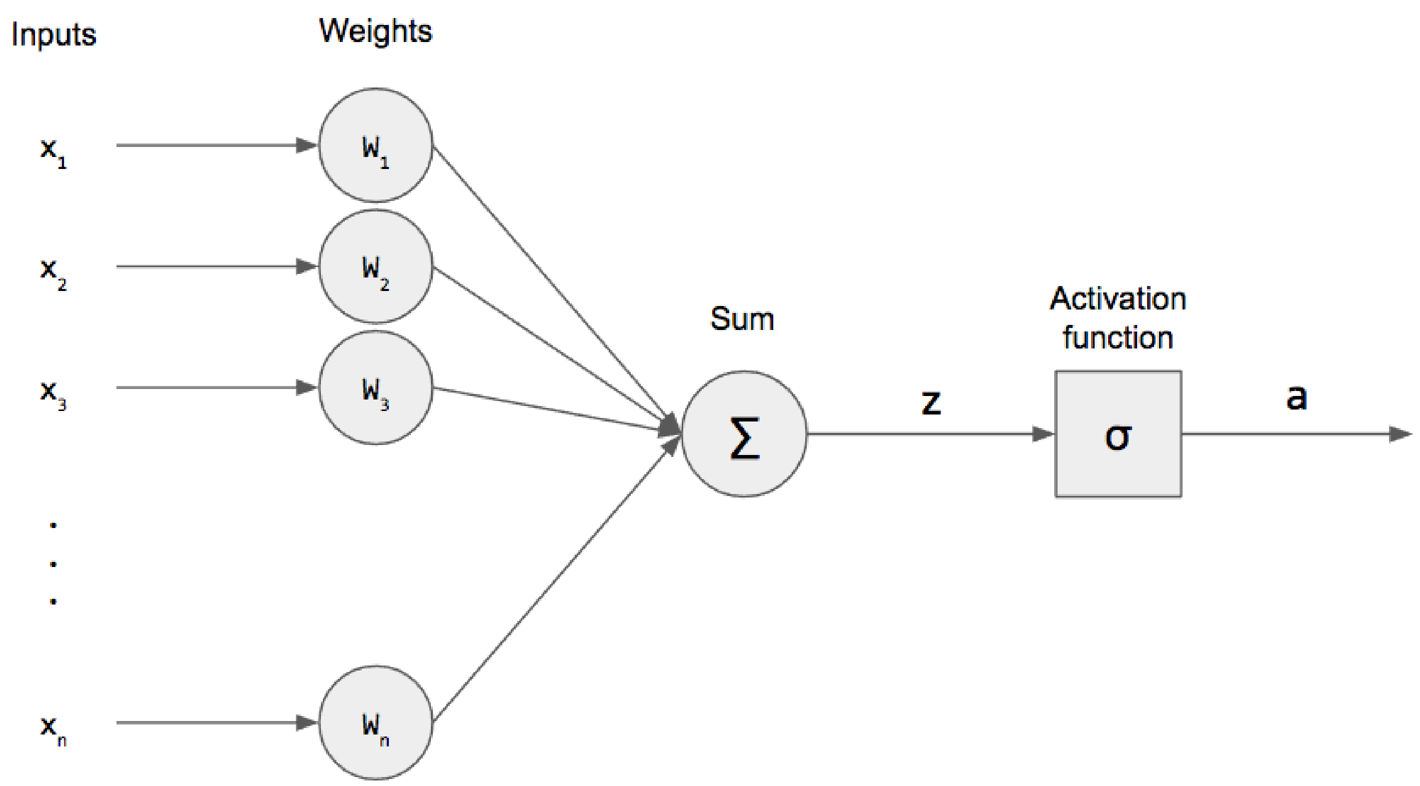
Modelos de Classificação
Support Vector Machine (SVM)

- Kernel Trick



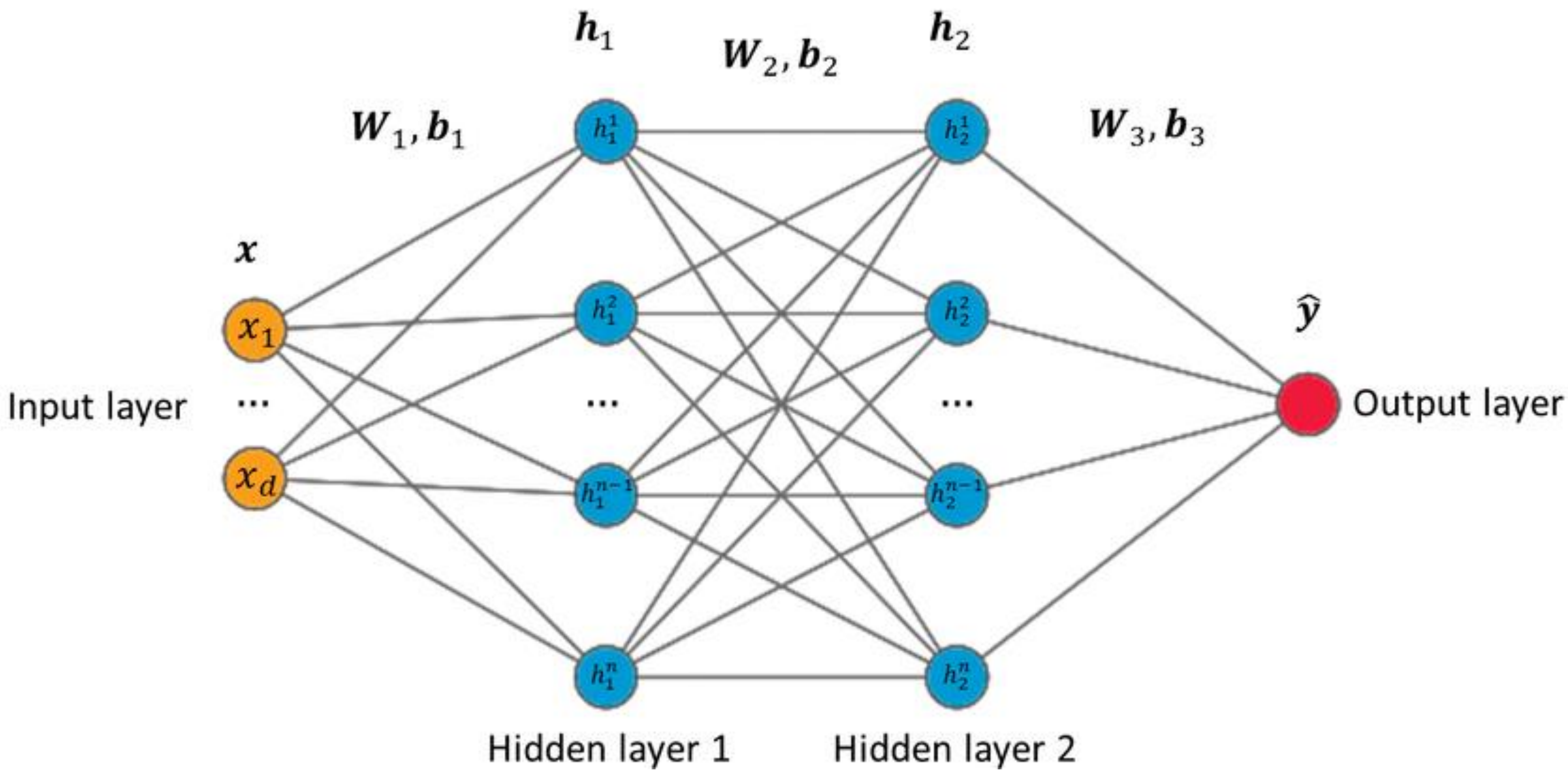
Modelos de Classificação
Multi-Layer Perceptron

- Perceptron



Modelos de Classificação
Multi-Layer Perceptron

- Multi-Layer Perceptron (MLP)

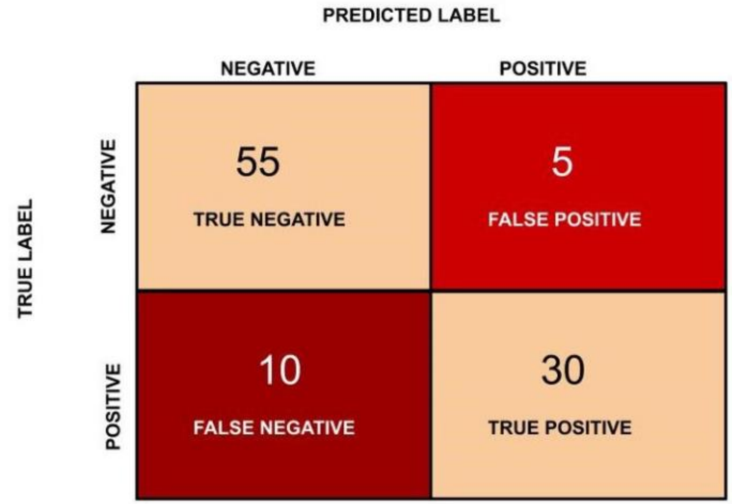


Métricas de Desempenho

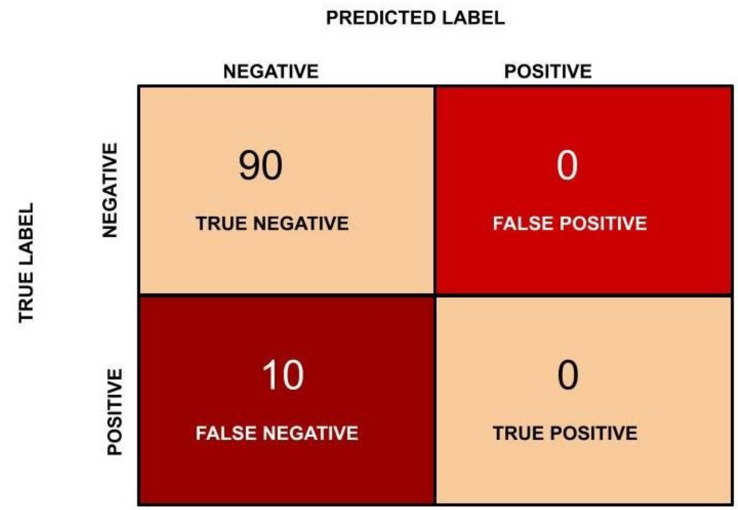
- Accuracy:
  - **Instâncias corretamente classificadas** sobre o total de instâncias

$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN}$$

  - (55 + 30)/(55 + 5 + 30 + 10) = 0.850



- Qual o problema com *Accuracy*?
  - Dados desbalanceados
    - Acc: 90% (90/100)
    - Error TP: 100% (10/10)

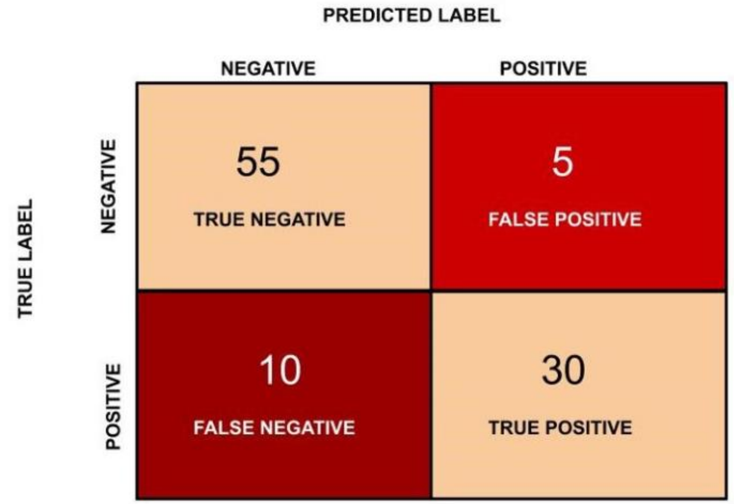


Métricas de Desempenho

- Precisão:
  - **Instâncias positivas classificadas corretamente** sobre o total de instâncias **classificadas** como positivas

$$Precision = \frac{TP}{TP + FP}$$

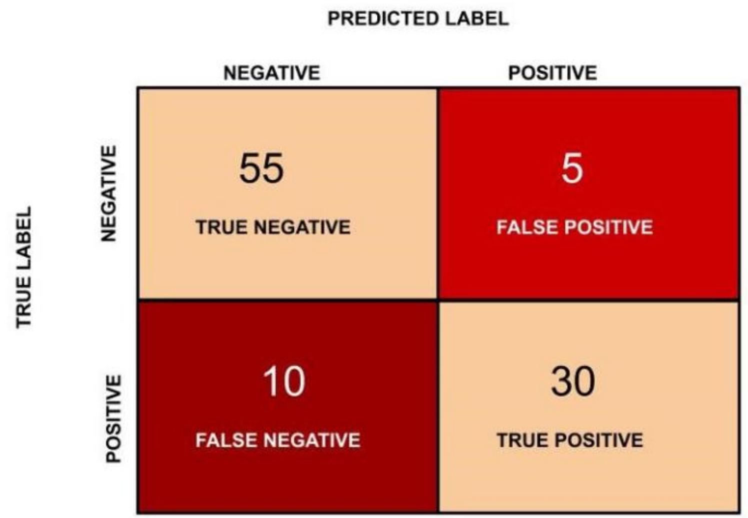
  - 30/(30+ 5) = 0.857



- Recall
  - **Instâncias positivas classificadas corretamente** sobre o **total de instâncias positivas** (A.K.A Sensitivity or TP Rate)

$$Recall = \frac{TP}{TP + FN}$$

  - 30/(30+ 10) = 0.750

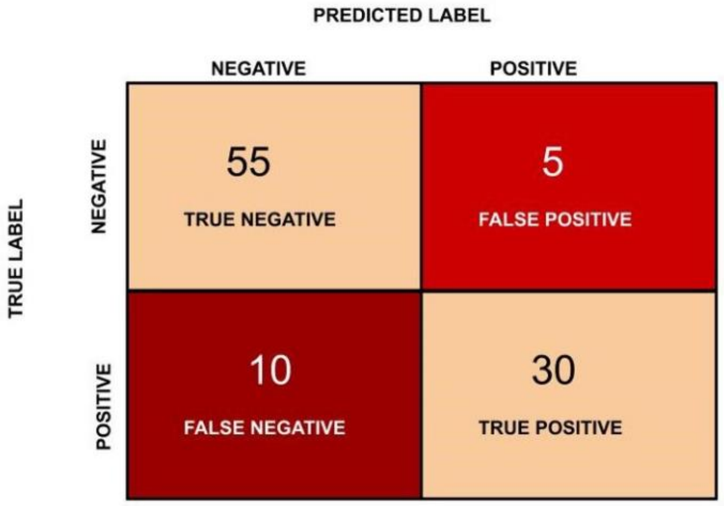


Métricas de Desempenho

- F1-SCORE:
  - Média Harmonica(*) entre precisão e recall

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

  - 2* (0.857 * 0.75)/(0.857 + 0.75) = 0.799



- Discussão
  - Accuracy: 0.850
  - F1-Score: 0.799
    - Precision: 0.857
    - Recall: 0.750

(*) The harmonic mean is a method that gives less weightage to larger single values and more weightage to smaller values

Codificação

- Siga o [LINK]